Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias

Year: 2023 | *Citations: 308* | *Authors: Yue Yu, Yuchen Zhuang, Jieyu Zhang*

Abstract

Large language models (LLMs) have been recently leveraged as training data generators for various natural language processing (NLP) tasks. While previous research has explored different approaches to training models using generated data, they generally rely on simple class-conditional prompts, which may limit the diversity of the generated data and inherit systematic biases of LLM. Thus, we investigate training data generation with diversely attributed prompts (e.g., specifying attributes like length and style), which have the potential to yield diverse and attributed generated data. Our investigation focuses on datasets with high cardinality and diverse domains, wherein we demonstrate that attributed prompts outperform simple class-conditional prompts in terms of the resulting model's performance. Additionally, we present a comprehensive empirical study on data generation encompassing vital aspects like bias, diversity, and efficiency, and highlight three key observations: firstly, synthetic datasets generated by simple prompts exhibit significant biases, such as regional bias; secondly, attribute diversity plays a pivotal role in enhancing model performance; lastly, attributed prompts achieve the performance of simple class-conditional prompts while utilizing only 5\% of the querying cost of ChatGPT associated with the latter. The data and code are available on \url{https://github.com/yueyu1030/AttrPrompt}.